Minimum Bayes-Risk Word Alignments of Bilingual Texts

July 7, 2002

Shankar Kumar and Bill Byrne

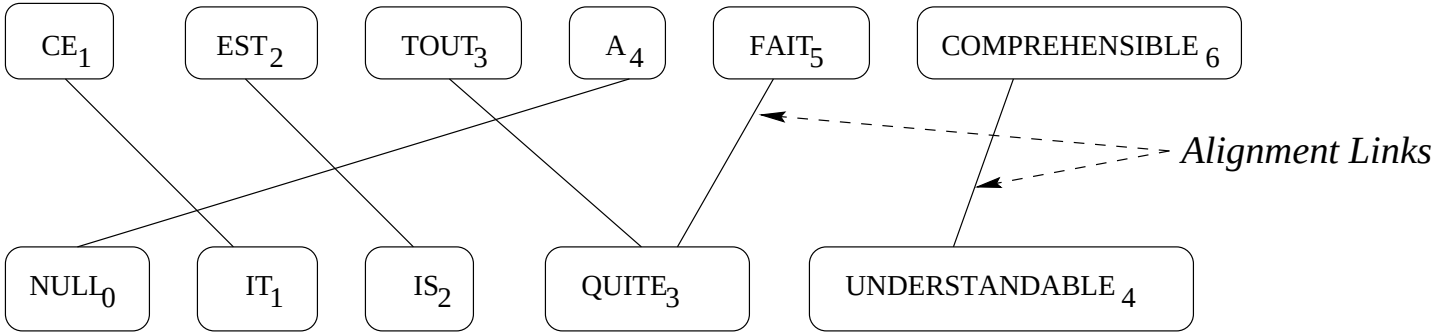
Center for Language and Speech Processing
Department of Electrical and Computer Engineering
The Johns Hopkins University, Baltimore, MD, U.S.A.

Minimum Bayes-Risk Word Alignment of Bitexts

- Word alignments of Bitexts are useful for:
  - Machine Translation
  - Dictionary Induction
  - Projecting Linguistic Structures such as Parse-Trees and POS Tags across Languages
- We introduce *Alignment loss functions* to measure alignment quality
  - Different loss functions capture different features of alignments
  - Will show that loss functions can use information from word-to-word links, parse-trees and POS tags
- We optimize alignment quality by generating *Minimum Bayes-Risk (MBR) Alignments* under each loss function
  - This requires an explicit loss function that can be computed between any two alignments
  - Will show performance gains by tuning alignment to the evaluation criterion

Word-to-Word Bitext Alignment

- A Candidate Alignment



- (e_0^l, f_1^m) : An English-French Sentence Pair
- *Alignment Links*: $b = (i, j)$: f_i linked to e_j
- Alignment is defined by a *Link Set* $B = \{b_1, b_2, ..., b_m\}$
- Some links are NULL links e.g. $A_4 - \text{NULL}_0$

Bitext Alignment Loss Functions

- Given a candidate alignment B' and the reference alignment B', $L(B, B')$ is the *loss function* that measures B' wrt B.
- Alignment Error Rate (AER) - Och and Ney (2000)
  - AER measures fraction of non-NULL links by which B' differs from B.
  - AER ignores all links to the NULL word
    - Define new linksets $(\bar{B}, \bar{B}')$ by removing all NULL links
    - Loss function defined on $\bar{B}$ and $\bar{B}'$
  - Reference Alignments are created by human experts
    - The experts identify which links are unambiguous
    - These unambiguous links are called sure links: $S \subseteq \bar{B}$
  - AER is defined wrt human generated alignments

$$AER(S, B; B') = 1 - \frac{|\bar{B}' \cap S| + |\bar{B}' \cap \bar{B}|}{|\bar{B}'| + |S|}$$

- For technical reasons, AER is an inconvenient loss function for MBR alignment

Loss Functions for MBR Alignment: Alignment Error

- Derived from AER

$$\begin{aligned} L_{AE}(B, B') &= |\bar{B}| + |\bar{B}'| - 2|\bar{B} \cap \bar{B}'| \\ &= |\bar{B}| + |\bar{B}'| - 2 \sum_{b \in \bar{B}} \sum_{b' \in \bar{B}'} \delta_b(b') \end{aligned}$$

- Measures # of links by which B' differs from B.
- Unlike AER, L_{AE} doesn't distinguish between sure and ambiguous links
- If $B' : L_{AE}(B, B') = 0 \Rightarrow \text{AER}(B, B') = 0$
- L_{AE} has limitations : it is only sensitive to link identities
  - All links that are incorrect are penalized equally (Similar to ASR Word Error Rate)
  - Suppose in any 2 candidate alignments, a specific French verb is aligned to an incorrect English verb in one alignment and an incorrect English noun in the other alignment, both these errors are penalized equally

Loss Functions for MBR Alignment: Generalized Alignment Error

- Extend L_{AE} loss function so that it can incorporate linguistic features

$$L_{GAE}(B, B') = 2 \sum_{b \in B} \sum_{b' \in B'} \delta_i(i') d_{ijj'}$$

$$b = (i, j), b' = (i', j')$$

- Word-to-Word Distance Measure $d_{ijj'} = D((j, e_j), (j', e_{j'}); f_i)$
  - Sensitive to both the English word identity and its position in e
- L_{GAE} can be almost reduced to L_{AE} (except in its treatment of NULL)

Examples of Word-to-Word Distance Measures for L_{GAE}

Suppose we have analyzed the English sentence. We would like L_{GAE} to be a function of the analyses

- Parse Tree Syntactic Distance: L_{PTS}

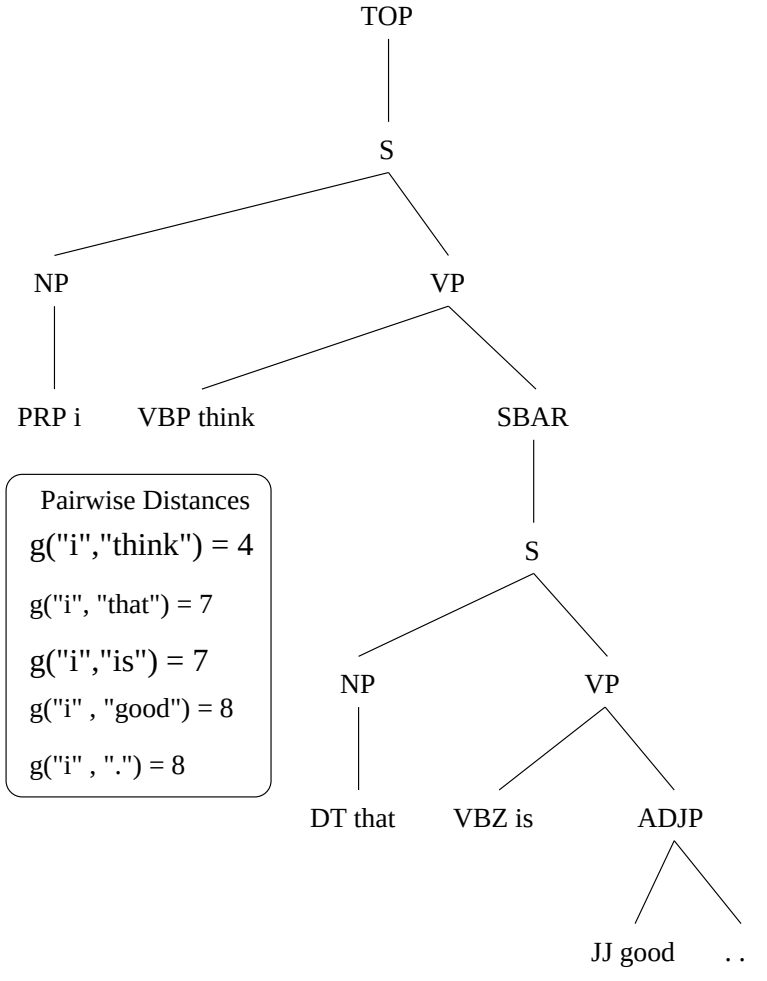
$$d_{ijj'} = g(N(e_j), N(e_{j'}))$$

- Part-of-Speech Tag Distance: L_{POS}

$$d_{ijj'} = \begin{cases} 0 & \text{POS}(e_j) = \text{POS}(e_{j'}) \\ 1 & otherwise. \end{cases}$$

- Automatic Word Class Distance: L_{AWC}

$$d_{ijj'} = \begin{cases} 0 & C(e_j) = C(e_{j'}) \\ 1 & otherwise. \end{cases}$$



Minimum Bayes-Risk Decoding for Automatic Word Alignment

- For the sentence pair (e, f) : $B' = \Delta(e, f)$
  - B' is the alignment produced by the decoder Δ
- Goal: Find the decoder $\Delta(e, f)$ that minimizes the Bayes-Risk

$$E_{P(B|f,e)}[L(B, \Delta(e, f))]$$

- $P(B|f, e)$: True Distribution of "human quality" alignments as found in actual word-aligned bitext.
  - Approximated by IBM-3 models.
- MBR decoder has the following well-known form (e.g. Goel and Byrne 2000)

$$\hat{B} = \underset{B' \in \mathcal{B}}{\operatorname{argmin}} \sum_{B \in \mathcal{B}} L(B, B') P(B|f, e)$$

- $\mathcal{B}$ is the set of all alignments of (e, f)
  - We approximate $\mathcal{B}$ by the alignment lattice: A set of the most likely word alignments of (e, f)
  - Each Lattice transition is an alignment link b & a lattice path defines a link set B

Derivation of MBR Alignment Under L_{AE} and L_{GAE}

$$\hat{B} = \underset{B' \in \mathcal{B}}{\operatorname{argmin}} \sum_{B \in \mathcal{B}} L(B, B') P(B|f, e).$$

Under L_{AE}, MBR decoder reduces to:

$$\hat{B} = \underset{B' \in \mathcal{B}}{\operatorname{argmin}} \sum_{b' \in B'} y_{b'}$$

$$y_{b'} = \begin{cases} 1 - 2P(b'|f, e) & j' \neq 0 \\ 0 & j' = 0. \end{cases}$$

Under L_{GAE}, MBR decoder reduces to:

$$\hat{B} = \underset{B' \in \mathcal{B}}{\operatorname{argmin}} \sum_{b' \in B'} z_{b'}$$

$$z_{b'} = \sum_{t \in \mathcal{T}} \delta_i(i') d_{ijj'} P(t|f, e).$$

- In each case, the MBR alignment is the *Consensus Alignment*:
  It is the most similar to all the competing alignments in the lattice under the loss function
- *Lattice Transition Posterior Probability* $P(t|f, e)$ is the sum of posterior probabilities of all lattice paths that pass through transition t
- *Alignment Link Posterior Probability* $P(b|f, e)$ is the sum of Lattice transition posterior probabilities of all transitions in the lattice that contain link b

Word Alignment Experiments - Setup & Lattice Generation

- Training Setup
  - Modified IBM-3 : Baseline Model to approximate $P(B, f|e)$
    - Distortion Model has a uniform distribution over all $m!$ permutations (Knight et al. 1998)
  - Training Set: Canadian English-French Hansards - 50K sentence-pairs
  - GIZA++ toolkit for training IBM-3 (Setup similar to Och and Ney - 2000)
- Alignment Lattice Generation : FSM Implementation based on Knight et al. (1998) using AT&T toolkit (Mohri et al. 2001)
  For a sentence-pair (e, f)
  - Build a Acceptor E for e
  - Build a transducer for each sub-model of IBM-3
    - Word Fertility(M), NULL Fertility(N), Word Lexicon(L)
  - Build a Permutation Acceptor P for f
  - Composition: $\mathcal{B} = E \circ M \circ N \circ L \circ P$
  - Prune the Composition Output $\rightarrow$ Lattice!
- Test Set: 207 French-English sentence pairs from Hansards: f has at most 16 words.

Minumum Bayes-Risk Alignment Experiments

Evaluation: Measure error rates wrt reference alignments from Aachen (F.J. Och)

Alignment Evaluation Metrics

- Alignment Error Rate (AER)
- Generalized Alignment Error Rate (GAER)
  $GAER(B, B') = \frac{L_{GAE}(B,B')}{|B|+|B'|}$
  - L_{GAE} : L_{PTS}, L_{POS}, L_{AWC}
  - Error Rate wrt all links in Reference Alignment

		Generalized Alignment Error Rates		
Decoder	AER (%)	PTS (%)	POS (%)	AWC (%)
ML	18.13	29.39	51.36	54.58
AE	**14.87**	19.81	36.42	38.58
PTS	23.26	**14.45**	26.76	28.42
POS	28.60	15.70	**26.28**	29.48
AWC	24.71	14.92	26.83	**28.39**

MBR decoder tuned for a loss function
performs the best under the corresponding error rate

Conclusions and Future Work

- MBR allows bitext word alignment procedure to be tuned for specific loss functions
- Our MBR decoders are built on top of IBM-3 statistical MT models
  In general MBR decoding strategies can be applied using other MT model architectures
- Syntactic features from English parsers and POS taggers can be integrated into a statistical MT system via appropriate definition of the alignment loss functions
  - without retraining it from scratch
- Future Work
  - Construct MBR decoders based on loss functions sensitive to word alignment and parse-trees in both English and French
  - Improve French parsing accuracy using *MBR-aligned* bitexts under appropriate loss functions (Hwa et al. 2002)
  - Incorporate linguistic knowledge such as morphology and base noun-phrases into the MBR alignment framework via newer loss functions